

AUTOMATIC METHOD FOR SELECTION OF KEY PHRASE FROM DOCUMENT OF MACHINE-READABLE FORMAT TO PROCESSOR

Publication number: JP8305730

Publication date: 1996-11-22

Inventor: FURANSHIINU AARU CHIEN; SUTEIBUN BII
PATSUTSU; DANIERU SHII BUROTSUKII

Applicant: XEROX CORP

Classification:

- International: **G06F17/30; G06F17/30; (IPC1-7): G06F17/30**

- European: **G06F17/30T1E**

Application number: JP19960105786 19960425

Priority number(s): US19950432383 19950501

Also published as:



EP0741364 (A1)

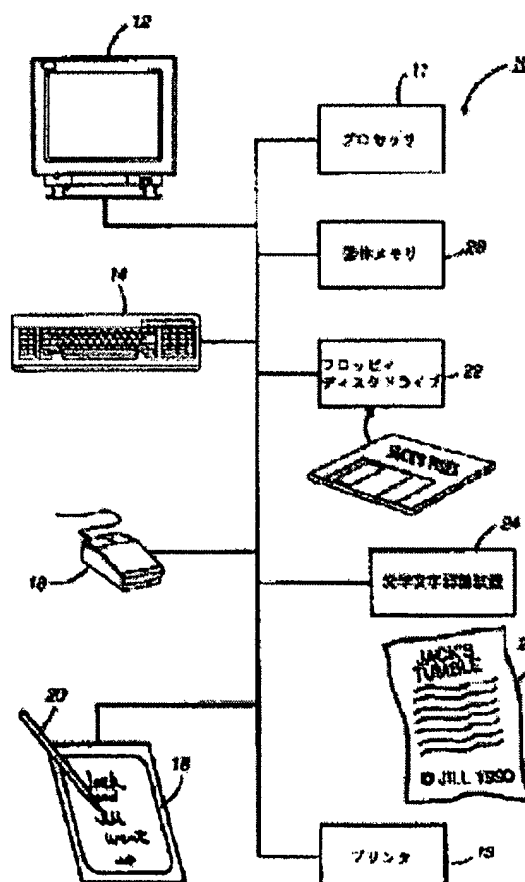
US5745602 (A1)

Report a data error here

Abstract of JP8305730

PROBLEM TO BE SOLVED: To automatically select a key phrase from a document in a machine readable format by generating many candidate phrases including more than two words from the document, and selecting the subset of the candidate phrases as a key phrase.

SOLUTION: A processor 11 judges whether or not the word is the acceptable starting part of a candidate phrase. After identifying the acceptable starting part of the candidate phrase, the processor 11 continues the construction of the candidate phrase from the most frequent term of a selection phrase until identifying the final word of the candidate phrase. Then, the processor examines the final word of the candidate phrase, and judge whether or not this is the acceptable final end of the candidate phrase. Next, the processor 11 judges whether or not the candidate phrase is sufficiently long, selects the word from the selection phrase, and judged whether or not the frequency of the selected word is high. As a result, the phrase is divided into non-overlapped partial phrases in the maximum length, and selected as a key phrase.



Data supplied from the esp@cenet database - Worldwide

(19) 日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11) 特許出願公開番号

特開平8-305730

(43) 公開日 平成8年(1996)11月22日

(51) Int.Cl. ⁶	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 17/30		9194-5L	G 0 6 F 15/401	3 1 0 A
		9194-5L	15/403	3 1 0 C

審査請求 未請求 請求項の数 3 O L (全 13 頁)

(21) 出願番号 特願平8-105786

(22) 出願日 平成8年(1996)4月25日

(31) 優先権主張番号 4 3 2 3 8 3

(32) 優先日 1995年5月1日

(33) 優先権主張国 米国 (U S)

(71) 出願人 590000798

ゼロックス コーポレイション

XEROX CORPORATION

アメリカ合衆国 ニューヨーク州 14644

ロチェスター ゼロックス スクエア

(番地なし)

(72) 発明者 フランシーヌ・アール・チェン

アメリカ合衆国 カリフォルニア州

94025 メンロパーク シャーマンアベニ

ュー 975

(74) 代理人 弁理士 小堀 益 (外1名)

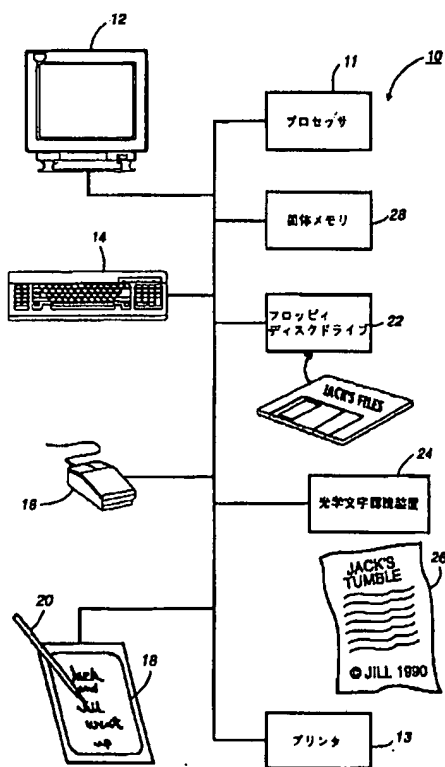
最終頁に続く

(54) 【発明の名称】 機械読み取り可能形式の文書からプロセッサに対してキーフレーズを選択する自動的方法

(57) 【要約】

【課題】 機械読み取り可能な形式で示すあらゆる文書からキーフレーズを選択できるようにすること。

【解決手段】 コンピュータシステム10で機械読み取り可能文書のテキストを複数語候補句に分割してキーフレーズを選択する。候補句は停止語を含まず、受容可能な語で始まり終了するものである。最後に最も頻度の高い候補句をキーフレーズとして選択する。



【特許請求の範囲】

【請求項1】 機械読み取り可能形式の文書からプロセッサに対してキーフレーズを選択する自動的方法であって、文書は第1の多数の語と第2の多数の文章を含み、文章内のいくつかの語は句を形成し、前記プロセッサは前記プロセッサに接続した記憶装置に記憶した命令を実行することで前記方法を実施するものであり、前記方法が、

a) 前記文書から各々の候補句が2つ以上の語を含む多数の候補句を生成するステップと、

b) 候補句の部分集合をキーフレーズとして選択するステップとからなる前記プロセッサで実施する方法。

【請求項2】 前記ステップa)が、

d) 前記第2の多数の文章の1つを現在文章として選択するステップと、

e) 前記選択文章のまだ検討していない語を選択語として選択するステップと、

f) 前記選択語がキーフレーズについて受容可能な開始部分かどうかを判定するステップと、

g) 前記選択語がキーフレーズについて受容可能な開始部分でない場合には、

1) 前記選択文章の全ての語を検討したかどうかを判定するステップと、

2) 前記選択文章の全ての語を検討していなければステップf)を繰り返すステップと、

3) 前記選択文章の全ての語を検討していればステップe)からステップf)を繰り返すステップと、

h) 選択語がキーフレーズについて受容可能な開始部分であるならば、

1) 前記選択語を現在句に加えるステップと、

2) 前記選択文書の全ての語を検討していなければ、前記選択文章のまだ検討していない語を選択語として選択してステップh 1)を繰り返すステップとからなる前記プロセッサで実施する請求項1記載の方法。

【請求項3】 前記ステップh 2)が、更に、

A) 前記選択文書の全ての語を検討していれば、

1) 現在句の最終語がキーフレーズについて受容可能な終端かどうかを判定するステップと、

i i) 現在句の最終語がキーフレーズについて受容可能な終端でなければ、現在句の最終語を除去してステップh 2 A 1)を繰り返すステップと、

i i i) 現在句の最終語がキーフレーズについて受容可能な終端であれば、現在句が2つ以上の語を含んでいるかどうかを判定するステップと、

i v) 現在句が2つ以上の語を含んでいれば、現在句を候補句リストに加えるステップとからなる請求項2記載の方法。

【発明の詳細な説明】

【0001】

【従来の技術】 キーワードリストにより読者は文書を読

まずにその文書の内容を判定することができる。文書のキーワードリストは自動的にあるいは人間の知性と労力を用いて文書を作成した後に作成することができる。しかし人間の労力を用いてキーワードリストを作成するにはコストが高くなる。これに対して、キーワードリストを生成する自動的手法を用いればコストは安くなる。

【0002】 文書のキーワードリストを自動的に生成する際には自然言語処理手法と統計的手法の両方が利用されてきた。自然言語処理は自然言語テキストを理解しようとするものであるので計算が膨大となる。統計的手法はテキストを理解する努力を行わないのでキーワードリストを速く生成することができる。1969年にキャロル(Carrol)及びロエロフ(Roeloffs)は「語頻度分析を用いたキーワードのコンピュータ選択」でキーワードを選択する方法を開示した。キャロル及びロエロフは各々の文書内ならびに文書コーパスにわたって語の相対的頻度に基づいてキーワードを選択した。しかし文書コーパスにわたる語頻度を使用する故にキャロル及びロエロフの方法は瞬時の結果を望む研究者や関連文書のコーパスを持たない研究者に取っては前処理無しには十分速いものとは言えない。

【0003】

【発明の実施の形態】 図1は本方法を実施するコンピュータシステム10をブロック図形式で示したものである。本方法はコンピュータシステム10の動作を変更して機械読み取り可能な形式で示すあらゆる文書からキーフレーズを選択できるようにするものである。要約すると、コンピュータシステム10で機械読み取り可能文書のテキストを複数語候補句に分割してキーフレーズを選択する。候補句は停止語を含まず、受容可能な語で始まり終了するものである。最後に最も頻度の高い候補句をキーフレーズとして選択する。以下にコンピュータシステム10を用いてキーフレーズを選択する2つの方法を詳細に説明する。

【0004】 A. キーフレーズ選択コンピュータシステム

本方法の詳細な説明を行う前に、コンピュータシステム10を考察する。コンピュータシステム10はコンピュータユーザに対して情報を視覚的に表示するモニタ12を有する。コンピュータシステム10は更にプリンタ13を通してコンピュータユーザに情報を出力する。コンピュータシステム10はコンピュータユーザに対して入力データに対する複数のルートを提供する。即ちキーボード14で入力することでコンピュータユーザはタイピングによりデータをコンピュータシステム10に入力することができる。またマウス16を動かすことで、モニタ12上に表示されたポインタを移動して表示されたアイコンを選択することができる。コンピュータユーザは更にスタイラスないしペン20でタブレット18に書き込むことで情報をコンピュータシステム10に入力でき

る。代わりにコンピュータユーザはフロッピーディスクなどの磁気媒体上に機械読み取り可能形式で記憶したデータをフロッピーディスクドライブ22にディスクを挿入することで入力することができる。光学文字認識装置(OCR装置)24によりコンピュータユーザはハードコピー文書26をコンピュータシステムに入力することもでき、そのOCR装置24は一般に情報交換用米国標準コード(ASCII)の符号化電子表示に変換する。

【0005】プロセッサ11はコンピュータシステム10の動作を制御、調整してコンピュータユーザのコマンド10を実行する。プロセッサ11は電子的にメモリに記憶した命令を実行することで各々のユーザコマンドに対応して判定して適切な処理を行う。一般にプロセッサ11の作動命令は固体メモリ28に記憶して命令に対する頻繁かつ高速アクセスを可能にしている。メモリ28を実現するのに利用できる半導体メモリには読み取り専用メモリ(ROM)、ランダムアクセスメモリ(RAM)、ダイナミックランダムアクセスメモリ(DRAM)、プログラマブル読み取り専用メモリ(PROM)、消去可能なプログラマブル読み取り専用メモリ(EPROM)20、そしてフラッシュメモリなどの電氣的に消去可能なプログラマブル読み取り専用メモリ(EEPROM)がある。

【0006】B. キーフレーズを選択する1つの方法

図2は機械読み取り可能文書からキーフレーズを選択するためプロセッサ11が実行する命令40を流れ図形式で示したものである。命令40は固体メモリ28ないしフロッピーディスクドライブ22に入れたフロッピーディスクに記憶することができる。命令40はLISPやC++を始めとするどの様なコンピュータ言語でも実現30できる。命令40の実行は機械読み取り可能文書の選択と入力で始める。所望により、命令40の実行前に、コンピュータユーザはデフォルト数から「P」と示す選択キーフレーズ数を変更することもできる。デフォルト数はどの様な値にも設定できる。1実施例では、デフォルト値は5キーフレーズに設定している。

【0007】プロセッサ11はステップ42に分岐してトークン化文書の選択に対応する。ここで用いるように、トークン化文書は文章境界及び語トークンが識別したものである。ステップ42中、プロセッサ11はトークン化文書を検査して複数語句を生成する。即ちプロセッサ11は各々の文章から2つ以上の語の非重複句を抽出する。句の各々の語が文書の主題に関連する意味を伝達するように、ステップ42中に生成される句から停止語は好適に除外される。停止語は文書の主題に関連した意味を殆ど伝えない代名詞、前置詞、限定詞、「to be」動詞などの語である。句から停止語を除外することはコンパクトなキーフレーズを生成し、ステップ42後のステップで必要な処理時間を削減できるという利点がある。プロセッサ11は各々の文章の各々の語トークン

ンを停止リストの語と比較することで停止語を除外する。プロセッサ11は文章内で停止語が出て来ればいつでも1つの句を終了し、別のものを始める。その結果、生成される句は隣接用語から構成される。ステップ42中の作用の結果、句のリストが生成される。句リストが完了すれば、プロセッサ11はステップ42からステップ43へ分岐する。

【0008】ステップ43中、プロセッサ11は引き続き使用するため、文書内の句リスト上の各々の語の頻度を判定する。ステップ42中に使用したトークナイザにより、プロセッサ11は、文書の各々の語をリストしその語が出現する各々の文書を識別する用語リストを調べることで句リスト上の各々の語の頻度を判定することができる。そのようなリストにより、プロセッサ11は句リスト上の各々の語に付いて文章IDの数を数えるだけでよい。その後、プロセッサ11はステップ43からステップ44へ分岐する。

【0009】ステップ44中、プロセッサ11は句リスト上の句から候補句を生成する。プロセッサ11は候補句を生成する間、要因の数を考察する。プロセッサ11は句の開始語と終端語を検査して候補句に適切かどうかを判定する。それにより後に選択するキーフレーズは妥当なものとなる。どの様にプロセッサ11がそれらのタスクを行うかは図3に関して後に詳細に述べる。

【0010】ステップ44中、プロセッサ11は更に句の各々の語を検討してその語の頻度が高いかどうかを判定する。句内の語の頻度と句の頻度自身は、最も頻度の高い句は文書の内容を最もよく示していると思われるので文書のキーワードを選択するのに使用する。プロセッサ11はある語が文章内で少なくとも最低回数出現すればそれを頻度の高いものと見なす。即ちプロセッサ11は文書内の語の出現回数をしきい値に対して比較し、出現回数がしきい値を超過すれば、その用語を頻度の高いものと見なす。頻度の低い用語は候補句から除外する。短い文書に付いてはしきい値は好適には1に設定される。その結果、少なくとも2回出現する用語だけが頻度の高いものと見なされる。長い文書に付いては、しきい値は高いものが望ましいであろう。候補句のリストを備え、プロセッサ11はステップ44からステップ46へ進む。

【0011】ステップ46でプロセッサ11は候補句のリストからP個のキーフレーズを選択するタスクを開始する。プロセッサ11は各々の候補句の文書内の出現回数に従って候補句リストを分類し始める。頻繁に出現する候補句は出現回数が少ない候補句よりも候補句の分類リストで高く位置づけする。候補句間の連携は語数ないし文字数に換算して測定した候補句長さ、どの句候補が最も頻繁な語を含むかに従って、あるいは最高平均語頻度に換算することを始め、いくつかの形で分類することができる。ステップ46の結果、プロセッサ11は候補

句を格付けしたリストを保持することになる。その後、プロセッサ11はステップ46からステップ48に進む。

【0012】ステップ48中、プロセッサ11は選択キーフレーズ数をゼロに設定して候補句リストからキーフレーズを選択する用意をする。それを行うとプロセッサ11はステップ50に進み、P個のキーフレーズが選択されたかどうかを判定する。選択数がPに等しくなければ、全てのキーフレーズがまだ選択されていないことになる。プロセッサ11はステップ50からステップ52に進んでこの状況に対応する。

【0013】プロセッサ11はステップ52で分類候補句の一番上の候補句を検査する。略してその句を「現在句」と称することにする。プロセッサ11はステップ52で現在句が既に選択したキーフレーズの1つの変形であるかどうかを判定する。ここで用いるように、変形とは別の句に関係しているが語順ないし語幹が異なるものである。例えば「テキスト分析システム」の可能な変形には「システムでテキストを分析」、「文書分析システム」及び「文書処理システム」がある。いくつかの自動テキスト処理手法を用いて変形分析を行うことができるので、ここでは変形分析を詳細に述べない。

【0014】変形分析に基づいてプロセッサ11はステップ52から2つの経路の1つを取る。分類候補句リストの一番上の候補句がキーフレーズの1つの変形でなければ、プロセッサ11はステップ52からステップ54に進む。ステップ54でプロセッサ11は現在候補句を分類候補句リストから除去し、現在候補句をキーフレーズリストに載せる。その後、プロセッサ11はステップ54からステップ56に進み、選択したキーフレーズの数に1だけ増分する。それを行うとプロセッサ11はステップ50に戻る。

【0015】ステップ52の変形分析で現在候補句がキーフレーズの1つの変形であることが分かればプロセッサ11の動作は異なったものとなる。それに対してプロセッサ11はステップ52からステップ58に分岐する。ステップ58中、プロセッサ11は現在候補句を分類候補句リストから除去し、妥当ならばキーフレーズリストを変更する。1実施例では、キーフレーズリストに既にある句が分類候補句リストからちょうど選択した句の部分句ならばそれを除去して置き換える。従って例えばプロセッサ11は、「南カリフォルニア海岸」よりも部分句の「南カリフォルニア」を除外する。どの変形を除外するかを判定する別の方法として句の最小頻度の変形を除外することなどをステップ58中に使用することができる。その後、プロセッサ11はステップ58からステップ50に戻る。

【0016】ステップ50に戻り、プロセッサ11はP個のキーフレーズを選択したかどうかを判定する。P個のキーフレーズを選択していなければ、プロセッサ11

は分類候補句リストからP個のキーフレーズが選択されるまでステップ52、54、56、58を通して分岐する。P個のキーフレーズを選択していれば、プロセッサ11はステップ50からステップ60に分岐し、文書のキーフレーズの選択を完了する。

【0017】B1. 候補句の生成

図3は句を最大長で受容可能に開始し終了する候補句に分割するステップ44の動作を詳細に例示したものである。要約すると、プロセッサ11は選択した句の各々の語を一時に1語ずつ検査してその語が頻度の高いものかどうかを判定する。ステップ44で生成した候補句は隣接し頻度の高い用語全体で構成されているので、句の長さや句内の頻度の低い用語の位置により、1つの句は複数の候補句を生成できたり全くなかったりする。選択した句の最初の頻度の高い語を識別すると、プロセッサ11はその語が候補句の受容可能な開始部分であるかどうかを判定する。候補句の受容可能な開始語を識別した後、プロセッサ11は候補句の最終語を識別するまで選択句の頻度の高い用語から候補句の構築を続ける。そしてプロセッサ11は候補句の最終語を検討してそれが候補句の受容可能な終端部分であるかどうかを判定する。そうでなければプロセッサ11は受容可能な終端語が見つかるまで候補句の最後から語を除去する。次にプロセッサ11は生じる候補句が十分長いものかどうかを判定する。プロセッサ11は候補句が十分な語数を含んでいればそれを記憶する。

【0018】以上の前提を想定して、ここで命令44の詳細な説明を助ける状況を考察する。第1に、ステップ42で生成した句のリストが「南太平洋会社は大きな影響を及ぼした」「4年後」「料金対無料」を含むものとする。第2に、更に文書内で2回以上出現する語に「南」「太平洋」「会社」「大きな」「影響力」「年」「後」「対」「無料」があるとする。第3に、不良開始リストに「対」が含まれるとする。最後に第4に、不良終端リストに「対」「後」が含まれると想定する。候補句の生成はステップ70で句リストから句の1つを選択することで始める。プロセッサ11はステップ70を通して第1の経路の「南太平洋会社が大きな影響力を及ぼした」を選択すると想定する。その後、プロセッサ11はステップ70からステップ72に分岐する。

【0019】ステップ72中、プロセッサ11は検査のため、選択した句の1つの語を選択する。好適には、選択した句の語の検査は左から右に順に進める。命令44が受容可能な開始部分の検査前に受容可能な終端部分を検査するように変更されていれば、選択句の検査は右から左に順に進めることもできる。プロセッサ11は選択句の語の検査を方向に関係なく進めるが、語は各々の生成された候補句が隣接用語で確実に構成されるように順に検査しなければならない。プロセッサ11は好適にはステップ72を通してその最初の経路の「南」を選択す

る。選択句から語を選択した後、プロセッサ11はステップ72からステップ74に分岐する。プロセッサ11はステップ74で、選択した語が頻度の高いものかどうかを判定する。プロセッサ11は選択した語の出現回数をしきい値と比較することでそれを行う。しきい値の値はキーフレーズが生成されている文書の長さ依存する設計上の選択である。1実施例では、しきい値は、各々の語の頻度が高いと見なすためには少なくとも2回出現しなければならないように1に設定する。

【0020】ステップ74の結果、句は最大長の非重複部分句に分割される。従って例えば「ニューメキシコ境界線」という句は、「ニューメキシコ」「メキシコ境界線」という部分句ではなく「ニューメキシコ境界線」という候補句だけを生成する。最大長の候補句だけを使用することで偽候補句を生成することがあるが、それらの候補句はその出現頻度が低い故にキーフレーズとして選択される可能性は低い。対照的に、最大長候補句から生成される部分句は、その語数が少ない故に頻繁に出現する可能性が高く、キーフレーズとして除外される可能性は低い。その結果、最大長候補句の部分句を用いて妥当なキーフレーズを生成するには、本方法を変更する必要がある。

【0021】「南」はここでの想定で頻度の高い語であるので、プロセッサ11はステップ74からステップ76に分岐して対応する。プロセッサ11は候補句の潜在的な開始語が識別されればステップ76に入る。プロセッサ11はステップ76で、選択語が候補句の受容可能な開始部分かどうかを判定する。プロセッサ11は選択語に付いて不良開始リストを探索することでそれを行う。不良開始リストにはキーフレーズに関して受容できない開始部分の語が含まれている。英語テキストの不良開始リストは簡潔なものになるが、偽ないし不適切と思われるキーフレーズを生成する可能性を削減するため疑わしいときは語を不良開始リストに含める傾向にある。非英語文書に関しては、異なる語は不良開始リストに含めるべきである。例えば「of」に相当するフランス語の「de」は、フランス語の名詞句は「noun de adjective」の形であるので、停止語に含めるべきではない。「de adjective」で始まるキーフレーズの生成を避けるため、「de」はフランス語不良開始リストに含めるべきである。

【0022】「南」という語はここで想定するキーフレーズに関して受容可能な開始部分を為しているので、プロセッサ11はステップ76からステップ78に分岐する。プロセッサ11はステップ78で新しい候補句を構築する過程を始めるが、それを現在候補句と称することにする。ステップ78中、プロセッサ11は選択語を現在候補句に追加する。それを行うと、プロセッサ11はステップ78からステップ80に進んで選択句から隣接する頻度の高い用語を現在候補句に追加し始める。プロ

セッサ11はステップ80で選択句がまだ検討すべき追加用語を含んでいるかどうかを判定する。プロセッサ11は選択句の全ての語をまだ検討していないのでステップ80からステップ81に分岐する。ステップ81でプロセッサ11は現在候補句に含める可能性のある選択句の次の語を選択する。選択句を想定し、左から右に順に進んで、プロセッサ11はステップ81で「太平洋」を選択する。その後、ステップ82でプロセッサ11は選択語は頻度の高いものであると判定する。それに対応して、プロセッサ11はステップ82からステップ78に戻る。プロセッサ11は「太平洋」をステップ78で現在候補句に追加し、その結果「南太平洋」となる。それを行うと、プロセッサ11はステップ80に進み、選択句にまだ検討していない語が含まれることを見いだす。

【0023】プロセッサ11はステップ81で「会社」を選択し、ステップ82に進む。プロセッサ11は選択語は文書内で2回以上出現するのでそれは頻度の高いものであることが分かる。その結果、プロセッサ11はステップ82からステップ78に分岐し、選択語を現在候補句に追加する。その結果、現在候補句は「南太平洋会社」となる。その後、プロセッサ11はステップ78からステップ80に分岐する。

【0024】ステップ80中、プロセッサ11は選択句にまだ検討していない語が含まれていることを見いだす。従ってステップ81でプロセッサ11は選択句の次の語の「及ぼした」を選択する。プロセッサ11は次のステップで「及ぼした」は選択文章内で頻度の高い語ではないことを見いだす。現在候補句の最も右側の語に隣接する頻度の低い語の出現によりそれは終端する。その結果、プロセッサ11は選択語やいずれのものも現在候補句に追加しない。プロセッサ11はこの状況にステップ82からステップ84に分岐することで対応する。

【0025】ステップ84でプロセッサ11は現在候補句の最終語が受容可能な終端部分かどうかをその語に関して不良終端リストを探索することで判定する。不良終端リスト上の語はキーフレーズを偽ないし不適切なものにする可能性のあるものである。不良開始リストにより、不良終端リストに載せた語は分析している自然言語の言語に依存して変化することがある。以前の想定では、「会社」は受容可能な終端部分となる。隣接し頻度の高い用語全体で構成され、受容可能に終了し始まる候補句を選択すると、プロセッサ11はステップ84からステップ88に進む。

【0026】プロセッサ11はステップ88で現在候補句が2つ以上の語を含むかどうかを判定する。単一語の句は、語に付いての言語的な情報なしにはキーフレーズリストで偽のものとして出現する可能性があるため、本方法ではキーフレーズとして選択しない。そのような言語的な情報を得るために時間を取るよりも、単一語の句は候補として受け入れない。現在候補句は2つ以上の

語を含んでいるので、プロセッサ11はステップ88からステップ90に進む。

【0027】プロセッサ11はステップ90で現在候補句をいままでリストした句候補と比較する。現在候補句は最初に生成されるので、ステップ90を通して第1の経路で、プロセッサ11は現在候補句は候補句のリストにないことを見いだす。それに対応してプロセッサ11はステップ94で現在候補句を候補句リストに追加し、その候補句に関してカウントを1に設定する。後にプロセッサ11は候補句に関連したカウントをキーフレーズを選択するのに使用する。その後、プロセッサ11はステップ94からステップ96に分岐して別の候補句の構築を始める。

【0028】別の候補句を構築する作業はステップ96で選択句の全ての語が検討されたかどうかを判定することで始める。選択句の「大きな影響力」という語がまだ検討されていないので、プロセッサ11はステップ96からステップ72に戻って対応して選択句のその検討を続行する。プロセッサ11はステップ72で「大きな」を選択語として選択する。その後、プロセッサ11はステップ74、76、78、80、81、82、84、88を通してちょうど説明したように分岐して選択句から「大きな影響力」という別の候補句を構築する。

【0029】最終的にプロセッサ11はステップ88からステップ90に分岐する。現在候補句が候補句のリストに既に含まれていれば、プロセッサ11はステップ90からステップ92に分岐する。ステップ92でプロセッサ11は現在候補カウントのカウントを1だけ増分する。それを行えば、プロセッサ11はステップ92からステップ96に分岐する。

【0030】ステップ96に戻ると、プロセッサ11は選択句の全ての語の検討がなされたことを見いだす。その結果、プロセッサ11はステップ96からステップ70に進む。ステップ96でプロセッサ11は「4年後」を選択句として選択する。引き続いてステップ72でプロセッサ11は「4」を選択語として指定する。プロセッサ11はステップ74中に「4」は選択した文書内で頻度の高い語でないことが分かる。それに対応してプロセッサ11はステップ74からステップ96に進む。プロセッサ11はステップ96で選択句にはまだ検討していない語が含まれていることを判定する。プロセッサ11はステップ96からステップ72に戻って選択句の次に語を選択する。プロセッサ11は「年」を選択語として選択して選択語は頻度の高いものであると判定する。その結果、プロセッサ11はステップ76に進み、ステップ76で「年」に関して不良開始リストを探索するが、それが見つからないと「年」は受容可能な開始部分であることになる。

【0031】プロセッサ11はステップ76からステップ78に分岐して現在候補句の構築を続行する。選択語

はステップ78で現在候補句に追加する。次のステップのステップ80で、プロセッサ11は選択句にまだ検討していない別の語が含まれているかどうかを判定する。そうであればステップ81でプロセッサ11は「後」を選択語として指定する。次にプロセッサ11はステップ82で「後」は選択文書内で頻度の高い語であることを見いだす。プロセッサ11はステップ78に分岐し選択語を現在候補句に追加して対応する。この動作の結果、現在候補句は「年後」になる。その後、プロセッサ11はステップ78からステップ80に分岐する。

【0032】プロセッサ11はステップ80で選択句が追加語を含むかどうかを判定することで追加語を現在候補句に追加できるかどうかを判定する。プロセッサ11は選択句の全ての語を検討し終ると、現在候補句に対して更に追加するものはなくなり、ステップ80からステップ84に進んで対応する。プロセッサ11はステップ84で「後」に関して不良終端リストを探索して現在候補句が受容可能に終了するかどうかを判定する。プロセッサ11はステップ84からステップ86に分岐して不良終端リストに「後」が見つかることに対応する。そのステップでプロセッサ11は現在候補句から最終語を除去して現在候補句を「年」とする。その後、プロセッサ11はステップ86からステップ84に戻り再び現在候補句の最終語を検討する。不良終端リストに「年」はないので、プロセッサ11はステップ86からステップ88に分岐して対応する。ステップ88ではプロセッサ11は現在候補句が複数句であるかどうかを判定する。現在候補句は1つの語しか含まないので、プロセッサ11は現在候補句を捨ててステップ88からステップ96に分岐する。

【0033】プロセッサ11はステップ96で現在候補句の全ての語は既に検討してしまったので別の句を選択して検討しなければならないことを見いだす。その結果、プロセッサ11はステップ98に進んでまだ検討していない別の句があることを見いだす。プロセッサ11はステップ70に戻り、「料金対無料」を選択する。続いてプロセッサ11は「料金」を選択して検討し、ステップ72からステップ74へ分岐する。

【0034】プロセッサ11はステップ74で「料金」は頻度の高い語ではないことを見いだす。それに対応してプロセッサ11はステップ72に戻って選択した句の次の語の「対」を選択する。プロセッサ11は「対」は選択文書内で2回以上出現するので頻度の高い語であると見なす。それに従ってプロセッサ11はステップ74からステップ76に分岐する。プロセッサ11はステップ76で選択語に関して不良開始リストを探索してそれをそこで発見する。それに対応してプロセッサ11はステップ76からステップ96に分岐する。選択句の全ての語をまだ検討していないので、プロセッサ11はステップ96からステップ72に戻る。プロセッサ11はス

ステップ72で別の語を選択してステップ74に進む。プロセッサ11はステップ74で選択した語の「無料」は選択文書内で頻度の高い用語であると判定する。更に次のステップで、プロセッサ11は選択語は受容可能な開始部分であると判定する。それに対応してプロセッサ11はステップ78へ分岐して前述したようにステップ78、80、94、88、96、98を実行する。プロセッサ11は全ての句を検討したことをステップ98で見いだすまで命令44の実行を続行する。それが為されると、プロセッサ11はステップ98からステップ100

【0035】C. キーフレーズを選択する別の方法

図4は機械読み取り可能な形の文書からキーフレーズを選択する別の命令40aを流れ図形式で示したものである。命令40aは固体メモリ28ないしフロッピディスクドライブ22に入れたフロッピディスクに記憶することができる。命令40aはLISP及びC++を含むどの様なコンピュータ言語でも実現することができる。

【0036】命令40aは命令40とは、プロセッサ11は命令40を用いて選択するように同一句をキーフレーズとして必ずしも選択しなくてもよいという点で異なる。命令40aは更にプロセッサ11がキーフレーズをより速く選択できるようにする点で命令40と異なる。命令40aによりプロセッサ11は文書から必要な情報を、命令40では2回のパスを必要とするのに対して、1回のパスで抽出できる。命令40aは命令40に比べてメモリの使用を増大してこの速度的な利点を達成する。それらの相違にも関わらず、命令40aは命令40と非常に似ている。この類似故に、図4ではステップ44aと45だけを例示し、命令40aはステップ42ないし46に相当するものは含んでいない。図4ではステップ48-60はキーフレーズを選択する両方法に関して本質的に同一であるのでそれらのステップを例示していない。その結果、ステップ48-60は命令40aの以下の説明では述べる必要がない。

【0037】プロセッサ11はステップ44aで命令40aの実行を開始する。ステップ44aでプロセッサ11は停止語及び受容可能な開始及び終端語を識別することで候補句表を生成する。ステップ44aでプロセッサ11は候補句に含まれる語が頻度の高いものかどうかを考察しない。

【0038】ステップ44aでどの様に候補句表が構築されるかの説明を始める前に、まず句表の内容を考察する。句表は句カウント及び総称形式表示と表面形式表示の各々の候補句の2つの表示方法を含む。それらの表示が全く異なれば、候補句の語の大文字使用に関して異なることになる。候補句の総称形式表示は候補句の小文字バージョンであるが、文書内ではそれは出現しない。プロセッサ11は候補句に関して総称形式を判定し句表内

でその総称形式表示を探索することで、総称形式表示を句表へのキーとして使用する。プロセッサ11が句表内で候補句の総称形式表示に遭遇すると、その候補句を句表に追加する必要はない。その代わり、プロセッサ11は総称形式に関連した句カウントを増分する。表面形式表示は実際に大文字にした候補句の出現の1つを示すものである。表面形式表示によりプロセッサ11は、コンピュータユーザに各々のキーフレーズを文書内で少なくとも1回実際に大文字にされたものとして提示できる。好適に表面形式表示は常に候補句の出現を最小の大文字で示す。

【0039】プロセッサ11は総称及び表面形式の両方の候補句を語ID列として表現する。各々の語IDは語の1つのASCII表示に対して一意的な整数である。その結果、同一語の異なる大文字化により、異なるASCII表示故に異なる語IDを有することになる。例えば「hate speech」及び「Hate speech」という句は異なるASCII表示と異なる語IDを有する。プロセッサ11は語IDを語ID表から得る。プロセッサ11は句表と同時にステップ44aで語ID表を生成する。ステップ44aで語を選択して検討する度に、プロセッサ11はその後のASCII表示に関して語ID表を探索する。語ID表に語のASCII表示が含まれなければ、プロセッサ11はその表示を語ID表に加え、一意的な整数を指定して語IDとして機能させる。プロセッサ11は他の有用な情報を語ID表に格納して句表の生成速度を速める。文書の分析を始める前に、プロセッサ11は語を停止、不良開始及び不良終端リストから表に追加し、その後に関連したフラグを設定して語表を初期化する。従って例えば「the」という停止語を語ID表に追加する場合には、「the」に関連した停止語フラグが設定される。それらのリストの語を語ID表に追加する結果、プロセッサ11は特定の語に関連した全ての情報を検索する際は語ID表だけを調べるだけでよい。

【0040】周知のハッシュ手法を用いてステップ44aの実行中に語ID表内及び句表の情報を効率的に探索できる。その結果、命令40aの実行中にそれらの表からどの様にプロセッサ11が情報を検索するかについて説明は行わない。

【0041】句表と語ID表の説明を備えて、候補句を生成する命令40aの詳細を例示する図5を考察する。命令40aは命令44に関して先述したのと実質的に同様の方法で候補句を生成する。その結果、以下の説明ではその先述の説明の知識を想定し、候補句を生成する2つの方法間の相違に焦点を当てる。命令44と44aの間の相違は、命令40aは候補句を停止語を含むトークン化文書を文書内の語の頻度の先験的な知識なしに候補句を生成するので生じる。その結果、命令40aは停止語であるが希な用語でないものを探索する。語の頻度を

使用せずに候補句を終了することで、命令44を用いて生成する候補句に比べて候補句の平均長と数の両方が増大する。

【0042】命令40aの実行はステップ70aで始める。ステップ70aで、プロセッサ11はステップ70のように句ではなく、ある文章を候補句の潜在的な源として選択する。その後ステップ72aで、プロセッサ11は選択語として選択文章の語の1つを指定する。ステップ72aからプロセッサ11はステップ74に進む。ステップ74でプロセッサ11は語ID表内の適切な項目を調べ、関連停止語フラグが設定されているかどうかを判定することで、選択語が停止語かどうかを判定する。そうであれば、選択語は句に関して受容可能な語ではなく、プロセッサ11はステップ96に進む。ステップ96、98の実行は、実質的に先述のものと同様に進められる。他方、選択語が停止語でなければ、プロセッサ11はステップ76に分岐する。

【0043】ステップ76から、候補句の生成は命令44に関して先述したものと実質的に同様の方法で3つの小さい相違点を有して進められる。第1には、プロセッサ11はステップ76、82a、86中にリストそれ自身を調べる代わりに、語ID表を調べて選択語が不良開始、不良終端ないし停止リストのいずれかにあるかどうかを判定する。プロセッサ11が語ID表内に選択語を見つけことができれば、ステップ76でプロセッサ11はその語の項目を表に加える。第2にステップ82a中に、プロセッサ11は図3のステップ82の場合のように文書内のそれらの頻度よりも、それらが停止語かどうかに基づいて現在句から語を排除する。

【0044】候補句の生成後、プロセッサ11はステップ90に進んで、句表をどの様に変更するかを判定する用意をする。プロセッサ11はこのタスクを語ID表を用いて現在候補句の総称形式及び表面形式表示を生成し、現在候補句の総称形式表示を句表に配置することで開始する。句表に総称形式表示があれば、現在候補句が句表内に既に含まれていることを示す。それに対してプロセッサ11はステップ92に進んで候補句に関連したカウントを増分する。ステップ92でプロセッサ11は更に候補句の現在表面形式表示が候補句の表面形式よりもより多くの大文字を含んでいれば、それを変更することができる。好適に、現在句が現在表面形式表示よりも

多くの大文字を含んでいる場合には、表面形式表示の変更は行わない。他方、プロセッサ11が現在候補の総称形式表示を見つけことができれば、プロセッサ11はステップ94に向けてステップ90を出る。ステップ94では、プロセッサ11は現在句の総称形式表示と表面形式表示の両方を句表に加え、関連句カウントを1に設定する。

【0045】ステップ44aで全ての可能な候補句を生成した後、プロセッサ11は図4に示すステップ45aに進む。ステップ45aでは、句表から候補句の部分集合を選択する。プロセッサ11はそれを文書内で最も頻繁に出現する候補句の部分集合を選択することで行う。ステップ45aで選択された句の数は出力するキーフレーズの数のPを越えるはずであるが、さもなくば設計上の選択となる。ステップ45aの実行後、キーフレーズの選択は先述のように進める。

【図面の簡単な説明】

【図1】 機械読み取り可能文書からキーフレーズを自動的に選択するコンピュータシステムを示す。

【図2】 機械読み取り可能文書からキーフレーズを選択する方法の流れ図である。

【図3】 句から候補句を生成する方法の流れ図である。

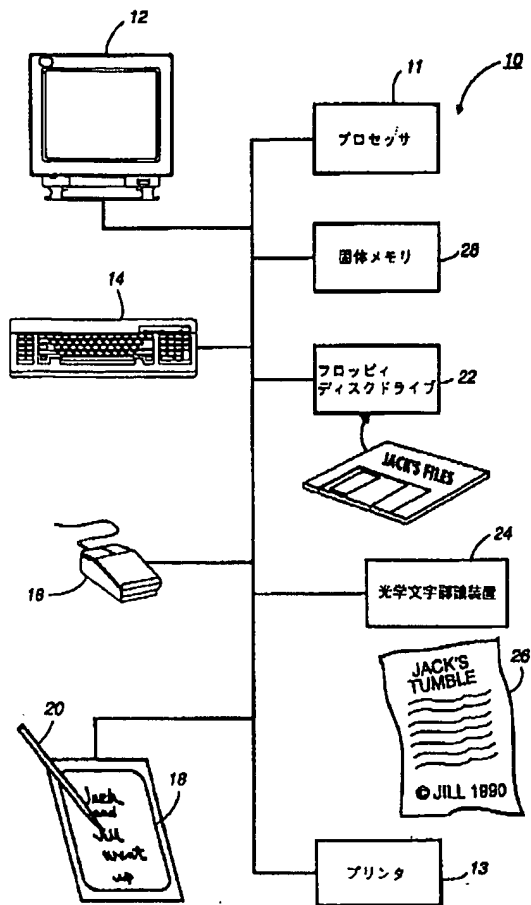
【図4】 キーフレーズを選択する別の方法の流れ図形式で示す。

【図5】 候補句を生成する別の方法の流れ図形式で示す。

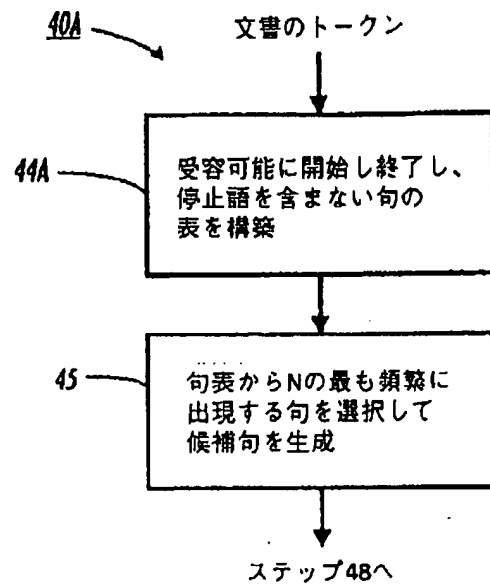
【符号の説明】

- 10 コンピュータシステム
- 11 プロセッサ
- 12 モニタ
- 13 プリンタ
- 14 キーボード
- 16 マウス
- 18 タブレット
- 20 スタイラスないしペン
- 22 フロッピーディスクドライブ
- 24 OCR装置
- 26 ハードコピー文書
- 28 固体メモリ

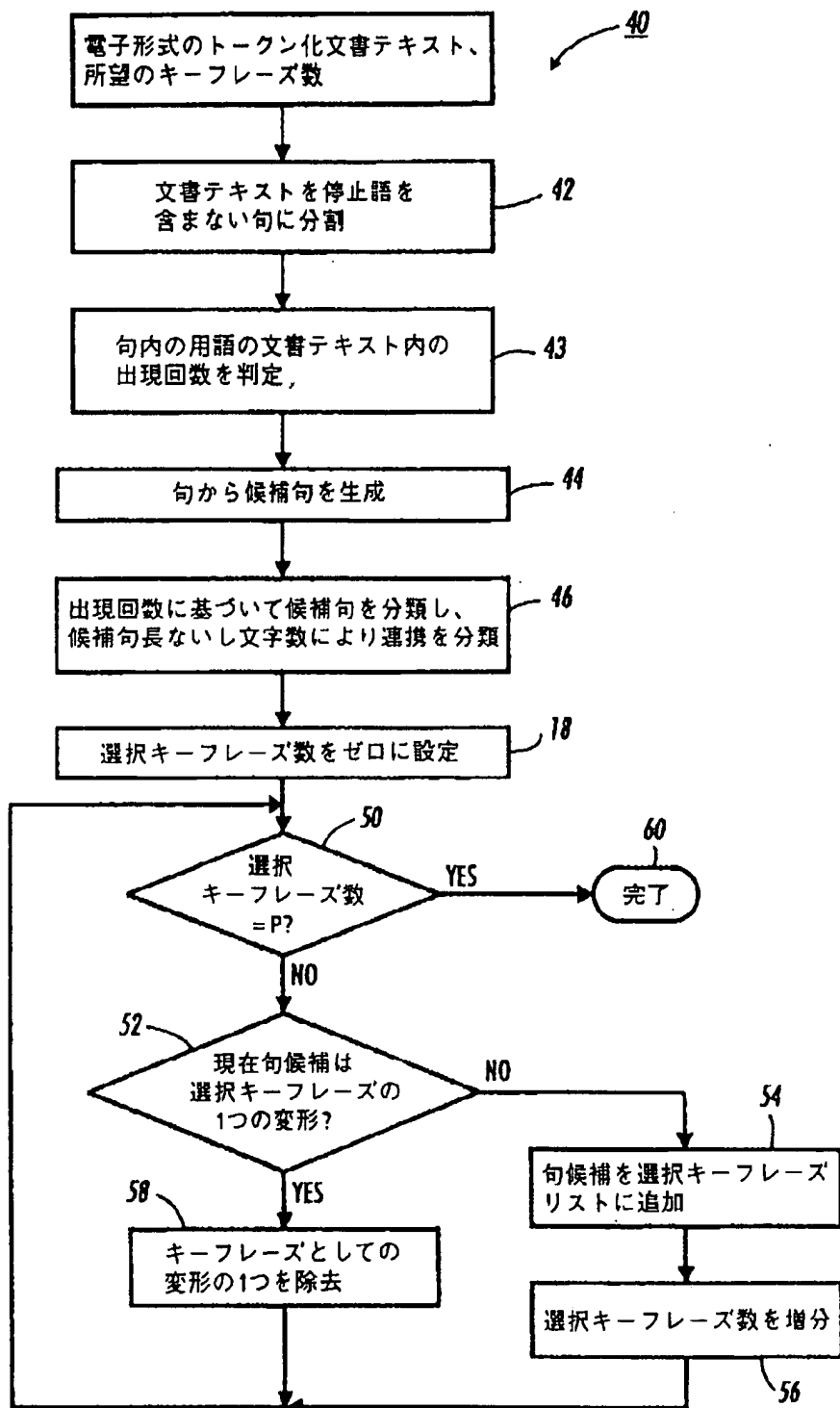
【図1】



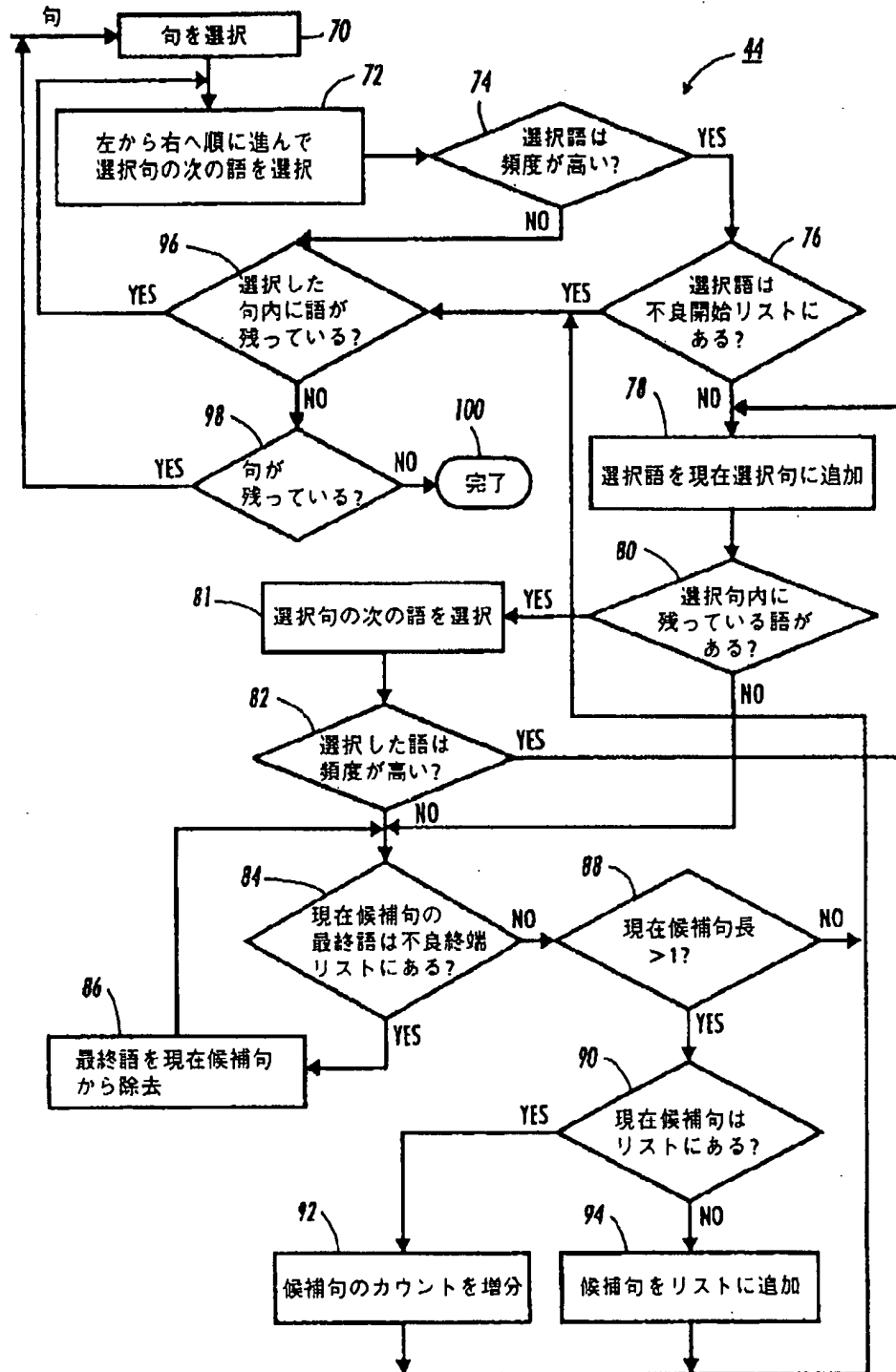
【図4】



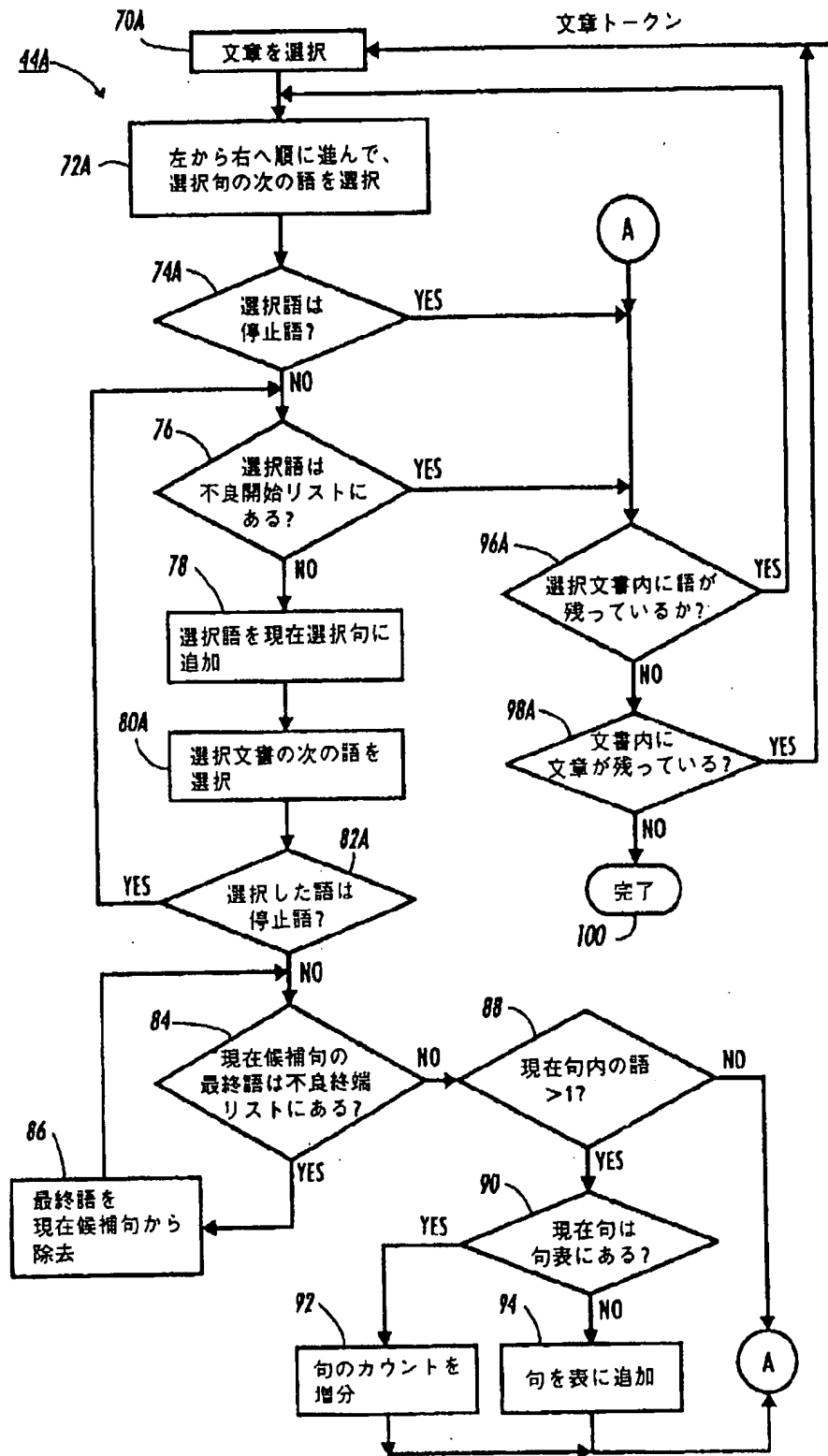
【図2】



【図3】



【図5】



フロントページの続き

(72)発明者 スティーブン・ビー・パッツ
アメリカ合衆国 カリフォルニア州
95051 サンタクララ ローズモンドドラ
イブ 351

(72)発明者 ダニエル・シイ・プロツキー
アメリカ合衆国 カリフォルニア州
94707 バークレイ コルサアベニュー
1162